SocialDict – 英文Webページのスマートな注釈・辞書引きシステム

SBM 研究会　2009/9/13

www.socialdict.com

（「補足」と題したページは発表中に説明できなかった部分の補足です．）

東京大学情報理工学系研究科

博士1年　江原　遥

自己紹介　niam

- 東京大学情報理工学系研究科　中川研究室

というところで,

- 自然言語処理
- 機械学習

の研究をしています.

Twitterやってます. はてなブックマーク, ダイアリーもやってます. Follow大歓迎！:

niam

mainの逆綴りの人.

今日の発表

発想の原点:「英単語を推薦したらどうなるか？」

SBMにおける推薦:

みんなが自分の気に入ったWebページをブックマーク
自分の興味のありそうなWebページを推薦してくれる

Webページ→(Webページ中の)英単語に

みんなが自分の知らない**英単語**をチェック
自分の知らなさそうな**英単語**を予測してくれる

真の狙い

- 英語って面倒だよね！

- 集合知で何とかできない？
- 「英語学習」の教材はたくさんあるが，今この場で読解を支援してくれるシステムを作りたかった.

SocialDict

名前はnokunoさんのSocial IMEをインスパイヤしました.

補足1　(1/2)

SBMとの関連がよく分からないという意見がありましたので，こういう狙いがあるよ，というものを1つ挙げます：

データを取るための手法としてみた場合：

- SBMはユーザの「Webページに対する興味」を計測する手法
- しかし，SBMでのブックマーク数が，必ずしもWebページの内容に対する興味を表していない場合もある．例えば，日本語のSBMでは英語のWebページのブックマーク数は少ない．
- 仮説：読めないページはブックマークされない

Webページのブクマ数＝
ユーザのWebページに対する興味*ユーザのWebページに対する読解力

そこで，ユーザの「Webページに対する読解力」を計測する手法として，SocialDictを提案．

補足1　(2/2)

ブクマ数＝Webページに対する興味　＊　Webページに対する読解力

- 通常，Webアプリケーションから各ユーザのWebページに対する読解力を取得することは難しいです．
- しかし，SocialDictを使えば，Webページ中の全単語について，各ユーザがそれらを知っている確率を計算できます．これを読解力の近似値として用いることが考えられます．
- ただ，このような目的（ブクマ数のモデル化）のためにSocialDictが役立つと主張するためには，学問的にはデータを溜めて，多くのユーザに関して上記の確率値を計算し，ブクマ数と実際につき合わせる必要があり，道は長いです．

語義注釈システム

Webページ中の分からない単語を辞書で引くことが可能な，読解支援システムの一種

既存の語義注釈システム

pop辞書 (2001)

pop-upで表示する

popIn (程, 2008)

なぞって選択した文字列に対して訳を埋め込む

政府は2005年から、温暖化防止対策(Global Warming Prevention)の一環として夏のビジネスで軽装を勧めるクールビズ運動を開始。官民を問わず定着しつつあった流れが、今回の選挙結果を受けて変わるのだろうか。

提案

予測機能を追加する:

クリックしたときに注釈を埋め込むのに加え，テキスト中のユーザが知らなさそうな単語（非既知語）を予測し，予め注釈をつける

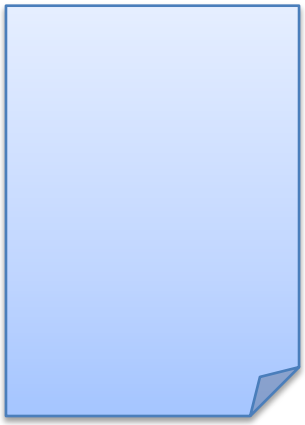
黄色：ユーザが知っていると判定された語
赤色：ユーザが知らないと判定された語

The BBC's Charles Scanlon said North Korea's diplomacy is following a familiar pattern - first belligerence: the walkouts and flaunting(みえを張る,誇示,(旗など)が翻る,得々と練り歩く,見せびらかす(show off),ひけらかす) of military muscle, followed by a return to diplomacy and demands for further concessions. (利権,特権,許容,免許,特許,譲歩,容認)

利点

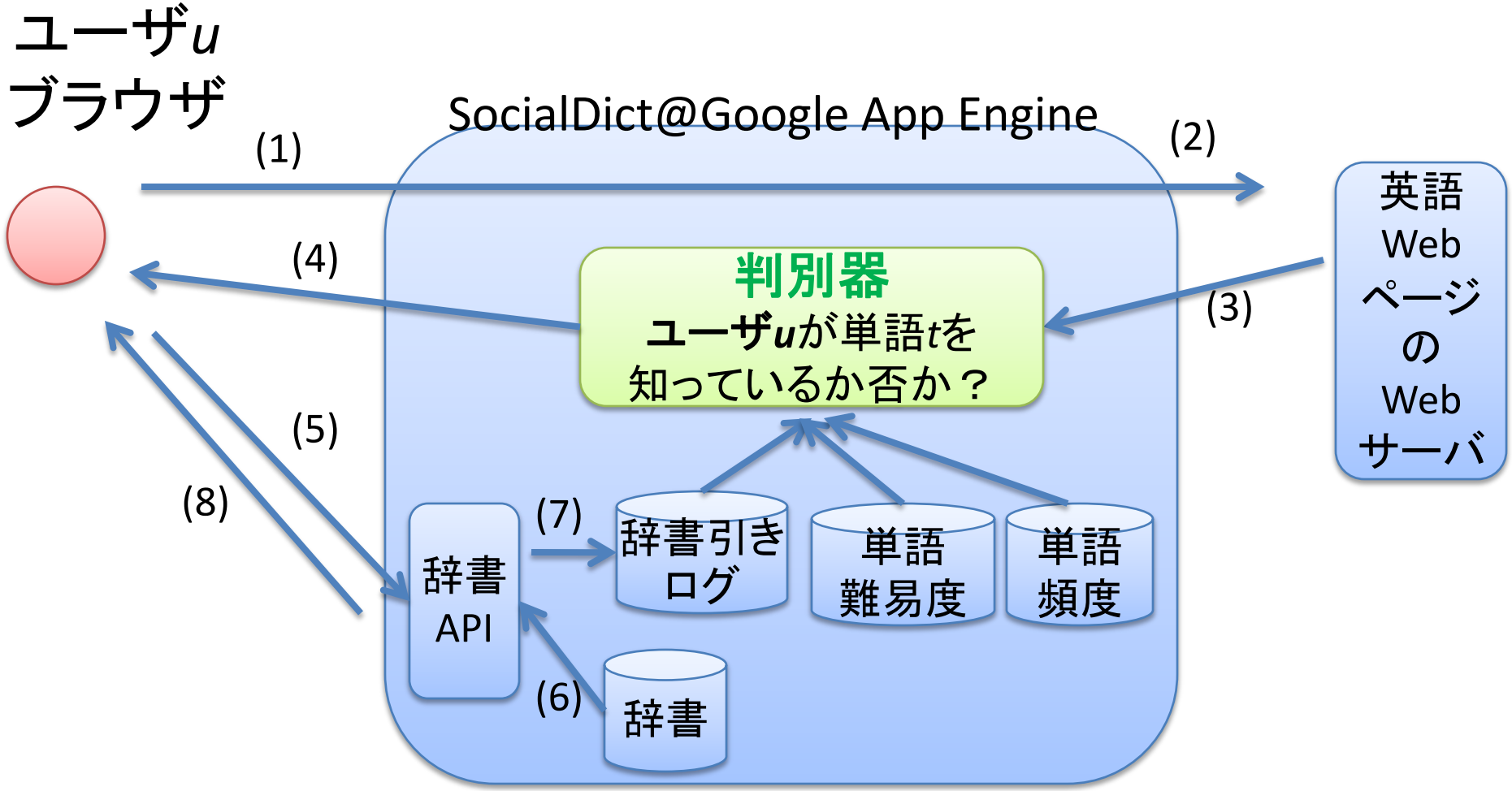
- クリックが不可能/難しい場合

印刷時（不可能）



PDAなど（難しい）

システムの概要



提案：辞書引きログを訓練データとしてユーザの語彙力を推定し，ユーザが知らなさそうな語彙を予め辞書で引いておく．

補足2

- 要するに，このシステムの構成で言いたかったのは，SBMでブックマークするときに，

http://b.hatena.ne.jp/*URL*

のようにしてブックマークできるのと同様に，

http://www.socialdict.com/*URL*

で英文の注釈＆予測が付けられるようにしましたよ，ということです．

他のSBMサービスと連携させて，同時にブックマークさせることも考えられます．

BBCニュース記事で比較

新規利用者

When the crew asked if there were doctors (博士(の資格),医師,《口語》医師を開業する,治療をする,治療を受ける,薬を飲む,混ぜ物を入れる,(報告・証拠などを)ごまかす,不正な変更を加える,(機械などを)修理する,博士号,医者,内科医(physician),《口語》修理屋,修繕屋,(俗)(船内,キャンプの)料理人,先生) on the flight, (フライト,逃亡,飛行,階段,逃避,便) Dr Julien Struyven, 72, a cardiologist (心臓内科医) and radiologist ( ) from ((原料・材料)〜から,から) Brussels, went to the cockpit and examined (試験する,尋問する,調べる,検査する) the pilot. (パイロット,水先案内人,を導く,を通す,試験的な,操縦する,導く)

"He was not alive," (生存している,活動状態の,生きている,活発な) Dr Struyven told (見分ける,話す,言う,教える,しかる,数える,表われる,分かる) AP. There was "no chance (偶然の,偶然,幸運,可能性,危険,機会,偶然である,予期せぬ / Chances are that 〜 : 恐らく〜であろう) at all" (すべての,すべての人[もの,こと],すべての人びと) of saving (1.節約する,救いの,節約(する),倹約(の),救済となる,倹約する,つましい,省力的な,2.救い,救援,救助,救済,省力,割引,値引率,3.〜以外は(except)) him, (彼を,彼に,彼) he said.

Dr Struyven said he suspected the pilot (パイロット,水先案内人,を導く,を通す,試験的な,操縦する,導く) had suffered (経験する,を被る,苦しむ,悩む,許す,受ける,損害を受ける) a cardiac arrest. (取り押さえる,逮捕する,つかまえる,引き止める,逮捕(する),止める) He said he used (利用,使う,利用する,(体,能力など)を働かす,使用,利用法) a defibrillator ( ) to try (1.〜を試験する,試みる,やってみる,試す,2.〜に苦難を与える,3.【法律】(事件を)審問する,審理する,裁判する,(人を)裁判にかける) and revive (生き返らせる,回復させる,元気にする,復興させる,蘇る,元気付く / The early dawn revived her.) the pilot, (パイロット,水先案内人,を導く,を通す,試験的な,操縦する,導く) but it was too (もまた,あまりに〜すぎて,あまりに) late.

Pilots are subject to rigorous medical checks (1.停止,妨害,検査,照合,小切手,反撃,勘定書,会計伝票,2.チェックする,急に止める,抑制する,調査する,一致する,3.(荷物などを)預ける) which increase in frequency with age. (年齢,時代)

自分（3000語訓練）

When the crew asked if there were doctors on the flight, Dr Julien Struyven, 72, a cardiologist (心臓内科医) and radiologist ( ) from Brussels, went to the cockpit and examined (試験する,尋問する,調べる,検査する) the pilot.

"He was not alive," (生存している,活動状態の,生きている,活発な) Dr Struyven told AP. There was "no chance at all" (すべての,すべての人[もの,こと],すべての人びと) of saving him, (彼を,彼に,彼) he said.

Dr Struyven said he suspected the pilot had suffered a cardiac arrest. He said he used a defibrillator ( ) to try and revive (生き返らせる,回復させる,元気にする,復興させる,蘇る,元気付く / The early dawn revived her.) the pilot, but it was too late.

Pilots are subject to rigorous medical checks which increase in frequency with age.

ピンク背景：非既知

黄色背景：既知

ピンク背景は色が濃いほど
知らない確率が高い

デモ

http://www.socialdict.com/

+見たいWebページのURL

辞書にはGENE95を使用

補足3 (1/2)

The BBC's Charles Scanlon said North Korea's diplomacy is following a familiar pattern - first belligerence: the walkouts and flaunting(みえを張る,誇示,(旗など)が翻る,得々と練り歩く,見せびらかす(show off),ひけらかす) of military muscle, followed by a return to diplomacy and demands for further concessions. (利権,特権,許容,免許,特許,譲歩,容認)

「知らないと判定された語をひっくり返せるのか(知っていることにできるのか)」という質問がありましたが，可能です.

上の例で，知らないと判定されたconcessionsを知っている場合，concessionsをクリックして語義注釈を引っ込めればOKです.

補足３ (2/2)

要するに，予測をクリックでひっくり返すと学習されます．

- 知っていると予測された単語（黄色）をクリック→実は知らなかったと学習される
- 知らないと予測された単語（赤色）をクリック→実は知っていたと学習される

The BBC's Charles Scanlon said North Korea's diplomacy is following a familiar pattern - first belligerence: the walkouts and flaunting(みえを張る,誇示,(旗など)が翻る,得々と練り歩く,見せびらかす(show off),ひけらかす) of military muscle, followed by a return to diplomacy and demands for further concessions. (利権,特権,許容,免許,特許,譲歩,容認)

Google App Engine (GAE)

- Webアプリケーション用クラウド環境.
- スケーラビリティを確保
  - ユーザが多数になっても対応可能
- ユーザ登録をGoogleアカウントで代用.
- Googleが2008年4月に提供を開始した

今はPythonで作ってます.

モデル

Item response theory (IRT)

IRT(項目反応理論):

TOEFLなどの既存の言語テストで利用されているモデルの総称．テスト結果から，ユーザの能力と項目（設問）の難易度を推定する．

Raschモデル:

IRTのうち最も基本的なモデル．蓄積される辞書引きログ(y_n,u_n,t_n)から，次のパラメータを計算する．

θ_u:ユーザuの語彙力，d_t:単語tの難しさ

ユーザ$u \in U$, 単語$t \in T$, $y \in \{0,1\}$

y=1:単語を知っている, y=0:単語を知らない

蓄積されるログ:

$(y_1, u_1, t_1), (y_2, u_2, t_2), \ldots, (y_N, u_N, t_N)$

Raschモデル

入力:

$(y_1, u_1, t_1), (y_2, u_2, t_2), \ldots, (y_N, u_N, t_N)$

出力:

θ_u:ユーザuの語彙力　　$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_u, \ldots, \theta_{|U|})$

d_t:単語tの難しさ　　$\mathbf{d} = (-d_1, \ldots, -d_t, \ldots, -d_{|T|})$

計算方法:

$$\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{d}} = \arg\max_{\boldsymbol{\theta}, \mathbf{d}} \prod_{n=1}^{N} P(y_n \mid u_n, t_n)$$

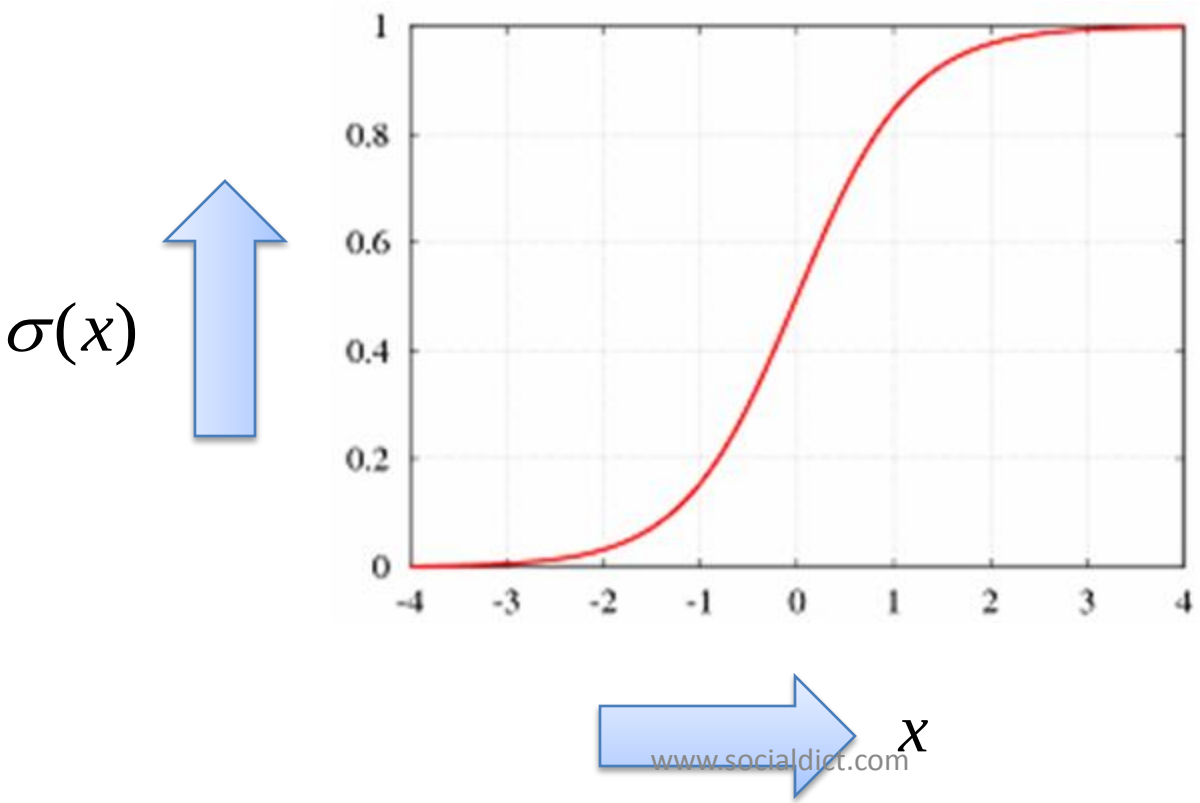
$$P(y_n = 1 \mid u_n, t_n) = \sigma(\theta_u - d_t)$$

ただし，$\sigma(x) = \dfrac{1}{1 + \exp(-x)}$とする.

シグモイド関数

$$\sigma(x) = \frac{1}{1+\exp(-x)}$$

任意の実数xを$(0, 1)$区間に写像して確率値とみなせるようにする



Raschモデルの強化

単純にRaschモデルを最尤推定で解く方法を強化し，本システムにあった方法を用いた.

- 予測性能の向上

  「d_t:単語tの難しさ」を補強する素性を追加

- Online学習

  Batch手法では，クリックごとにパラメータ推定をやり直さなければならず，逐次的にデータが蓄積される本システムには適さない.

素性を追加

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_u, \ldots, \theta_{|U|}), \mathbf{d} = (-d_1, \ldots, -d_t, \ldots, -d_{|T|})$$

$$\mathbf{e}_u = (0, \ldots, \underbrace{1}_{u}, 0 \ldots, 0)_{|U|} \qquad \mathbf{e}_t = (0, \ldots, \underbrace{1}_{t}, 0 \ldots, 0)_{|T|}$$

$$\mathbf{w}_{rasch} = \begin{pmatrix} \boldsymbol{\theta} & \mathbf{d} \end{pmatrix}^{\mathrm{T}}, \boldsymbol{\varphi}_{rasch}(u,t) = \begin{pmatrix} \mathbf{e}_u & \mathbf{e}_t \end{pmatrix}^{\mathrm{T}} \text{とすると,}$$

$$P(y_n = 1 | u_n, t_n) = \sigma\left(\theta_u - d_t\right)$$

$$= \sigma(\mathbf{w}_{rasch}{}^{\mathrm{T}} \boldsymbol{\varphi}_{rasch}(u,t)) \text{と表せる}$$

素性を追加

Rasch:

$$\mathbf{w}_{rasch} = \begin{pmatrix} \boldsymbol{\theta} & \mathbf{d} \end{pmatrix}^{\mathrm{T}}$$

$$\boldsymbol{\varphi}_{rasch}(u,t) = \begin{pmatrix} \mathbf{e}_u & \mathbf{e}_t \end{pmatrix}^{\mathrm{T}}$$

LOGRES:

（強化版）

$$\mathbf{w}_{LOGRES} = \begin{pmatrix} \boldsymbol{\theta} & \mathbf{d} & \mathbf{w}_a \end{pmatrix}^{\mathrm{T}}$$

$$\boldsymbol{\varphi}_{LOGRES}(u,t) = (\mathbf{e}_u \quad \mathbf{e}_t \quad \boldsymbol{\varphi}_a)^{\mathrm{T}}$$

φ_aに素性を追加することが可能．追加した素性：

- Google 1-gram
  約1兆ページのWebページ中の英単語の頻度．
- SVL12000
  人手で基本的な単語12,000語に対し，難易度を12段階でつけたもの．

Online学習を使用

Raschモデルの最尤推定の計算手法:

- Batch学習(L-BFGS, Nocealdal+, 89), (Trust region Newton method, Lin+, 08)
  最適解が得られる
  全データを何度も見て更新
- Online 学習(Stochastic Gradient Descent, SGD, Duda+, 2000)
  最適解を求めることは犠牲にする
  訓練データから逐次的に学習することが可能.
  →辞書引きログが逐次的に追加されていく今回のような実システムに適する。今回使用。

補足4　(1/2):2値を1本の式で表現

$$P(y_n = 1 | u, t; \mathbf{w}) = \sigma\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}\right) = \sigma\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}\right)^{y}$$

$$P(y_n = 0 | u; \mathbf{w}) = 1 - \sigma\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}\right) = \left(1 - \sigma\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}\right)\right)^{1-y}$$

より、

$$P(y | \boldsymbol{\varphi}; \mathbf{w}) = \sigma\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}\right)^{y}\left(1 - \sigma\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}\right)\right)^{1-y}$$

と表せる。

補足４　(2/2)：対数尤度関数を定義

結局、N本のデータを考え、次の対数尤度関数 $E(\mathbf{w})$を$\mathbf{w}$に関して最小化する問題に落とせる

$$E(\mathbf{w}) = -\log\left(\prod_{n=1}^{N} P(y_n \mid \boldsymbol{\varphi}_n; \mathbf{w})\right)$$

$$= -\sum_{n=1}^{N} \log\left(P(y_n \mid \boldsymbol{\varphi}_n; \mathbf{w})\right)$$

$$= -\sum_{n=1}^{N} \log\left(\sigma\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}_n\right)^{y_n}\left(1-\sigma\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}_n\right)\right)^{1-y_n}\right)$$

$$= -\sum_{n=1}^{N} y_n \log\left(\sigma\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}_n\right)\right) + (1-y_n)\log\left(1-\sigma\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\varphi}_n\right)\right)$$

SGD

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \prod_{n=1}^{N} P(y_n \mid u_n, t_n; \mathbf{w}) = \arg\min_{\mathbf{w}} E(\mathbf{w})$$

$$E(\mathbf{w}) = -\log\left(\prod_{n=1}^{N} P(y_n \mid u_n, t_n; \mathbf{w}) \right) \qquad , -\nabla E(\mathbf{w}) = \sum_{n=1}^{N} -\nabla E_n(\mathbf{w})$$

$$\eta_k = \frac{1}{\lambda(k + k_0)}$$

λ, k_0は定数
（ハイパーパラメータ）

SGD:
$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \eta_k \nabla E_n(\mathbf{w}^{(k)})$$

最急降下法:
$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \eta \sum_{n=1}^{N} \nabla E_n(\mathbf{w}^{(k)})$$

（比較）
ηは定数

評価

本研究の評価には語彙力の正解データが必要。

- 16人の被験者（東大修士院生）に対し
- １人12000語について

語彙力を測定した。

語彙力測定方法

自己申告式	折衷式	自己申告式
Dale (1965)	Paribakht+ (1993)	今回の設定
I never saw it before.	I have never seen this word.	見たこともない
		見たことがある気がする
I've heard of it, but I don't know what it means.	I have seen this word before, but I don't know what it means.	確実に見たことはあるが意味は知らない/覚えたことがあるが意味を忘れている
I recognize it in context.	I have seen this word before, and I *think* it means xxx.	意味を知っている気がする/意味が推測できる
I know it.	I know this word. It means xxx.	意味を確実に知っている
	I can use this word in a sentence.	

実験設定

実験目的:どの程度システムを使い込むと，どの程度の精度で予測可能なのかを測定する.

- 辞書引きログがN_0個蓄積されたところにユーザが1人新規にシステムを使い始め，N_1個の単語の既知/非既知が得られたと想定.
- この時，そのユーザのTestに含まれる単語のうち何%について既知/非既知を正解できたかを1人の精度とした.
- 12000語を

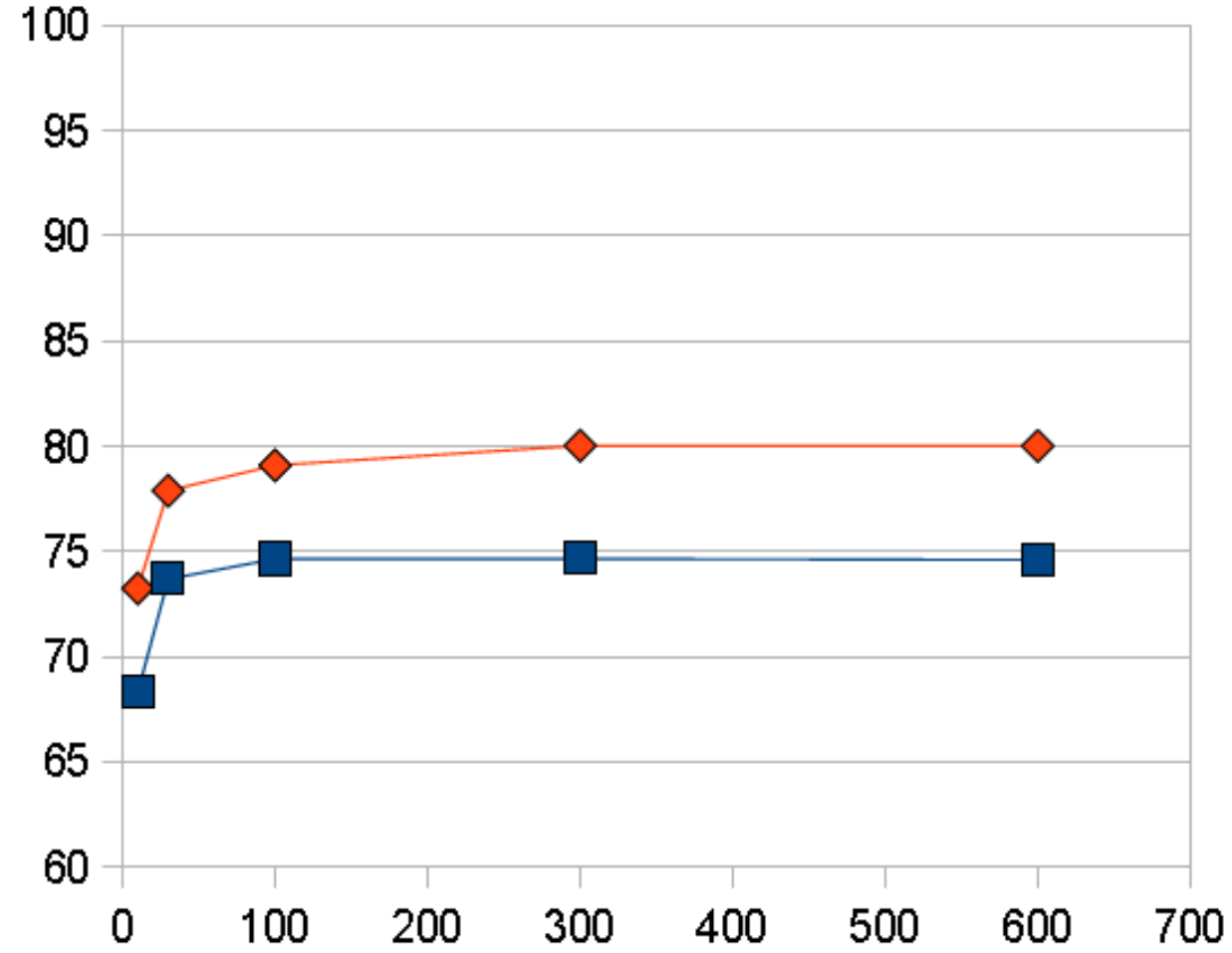
ログ（N_0)+10～600語(N_1)	Training
1400語	Development
9999語	Test

に分割し、16人の精度の平均値を計算。

- ログは，辞書引きログの代わりに，同じ(y_n,u_n,t_n)の形式のログが残るsmart.fm（旧iKnow)というシステムのログで代用

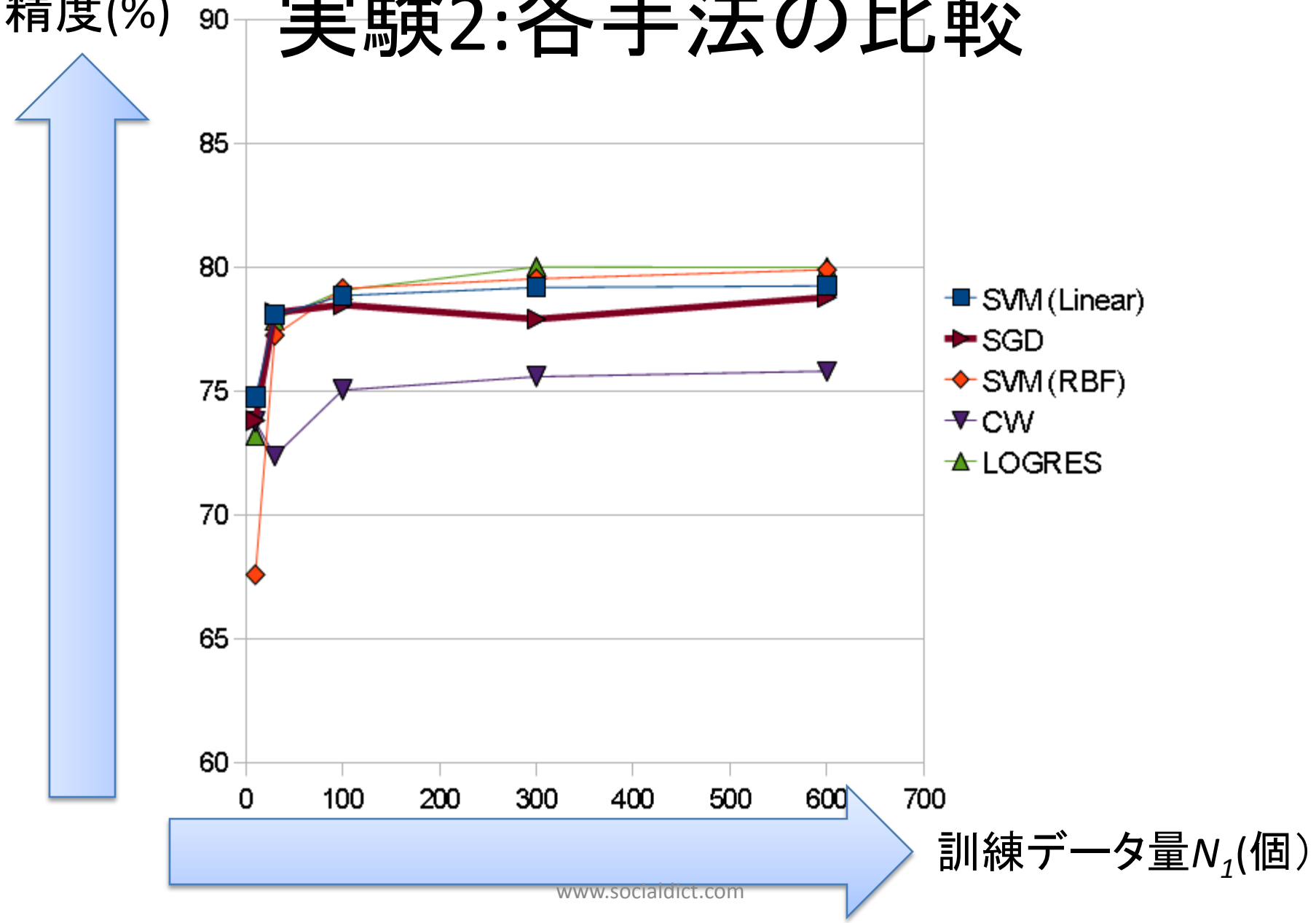
実験1:素性追加の効果

精度(%)



訓練データ量N_1(個)

実験2:各手法の比較



まとめ

目的：Webページ中の知らない英単語を予測（推薦）するシステムを作ってみた

結論：80%の精度で当たる

推薦アルゴリズムと数学的に同様な手法が，TOEFLなどの言語テストでも使われている.

（情報処理技術者試験でも使われているという噂）

- Log-linearモデル（対数線形モデル）
- ロジスティック回帰
- Raschモデル in 項目反応理論

- 学生の卒論，修論でこの手のサービスを作るなら，Google App Engineを使うのがベスト

TODO

ちょっとプリミティブすぎる・・・

- 優先的にやりたいこと：iKnowと連携
  - 知らない単語一覧をiKnowにインポート
- RSSフィード中のWebページの難易度を全部計算して，難易度順にソート

9月～10月中にはやりたい．

- PDFへの対応は，pdftotextなどを使えば楽だが，そうではないと難しい

データが溜まれば・・・

- データが溜まれば，面白いことが出来る
  - 例えば，ユーザのトピック（政治，スポーツなど）ごとの語彙力を表示したりだとか・・・

ので，みなさん使っていただければ幸いです．

研究としてやっているので，後２～３年はサポートします．特に，この冬の卒論・修論に向けて英語の文献をたくさん読む人の間で流行らせたいなぁ・・

ご清聴ありがとうございました